Anomaly Detection in E-Commerce Using Isolation Forest

Objective

To identify anomalous (outlier) customer behavior from an e-commerce dataset using the Isolation Forest algorithm and derive actionable business insights from the results.

Methodology

- **Dataset Size**: 3000 records (approx.) – post cleaning

- **Tech Stack**: Python, pandas, scikit-learn, seaborn, matplotlib

- **Preprocessing**:

  - Null value imputation using statistical and logical methods

  - Date parsed into month and year for seasonal trend analysis

  - Categorical variables encoded using one-hot encoding

- **Model Used**: Isolation Forest (tree-based anomaly detector)

  - n_estimators: 100

  - contamination: 0.05%

- **Output**: Binary label for outliers (1 = anomaly)

Findings from Outliers

- **Total Outliers Detected**: 149 (~5% of dataset)

- **Several outliers had**:

  - High annual income but **low purchase frequency and low spend**

  - Only **1 transaction** made

  - **High purchase** without being in a **loyalty program**

  - Concentration in certain **product categories**

Key Visuals

- **Purchase Amount vs Month**: Some months showed higher outlier activity (esp. January)

- **Category-wise Spend**: Toys and Home contributed most to anomalous purchases

Business Insights

- High-income, low-spend users - Upsell through personalized offers
- Non-members with high spend- Invite them into the loyalty member program
- January and December with high purchase amounts on outliers – make a check on fraud and review pricing
- Home and Toys with a high purchase amount on outliers – review product strategy

Recommended Actions

- Target outliers with relevant product bundles or engagement offers
- Investigate and validate data entry processes to reduce invalid entries
- Revisit loyalty program onboarding for high-spending but non-member customers
- Monitor high-risk categories more closely during peak months

Report Prepared by:

Karthikeyan S R